Available on the CERN CDS information server **CMS PAS JME-07-003**

CMS Physics Analysis Summary

Contact: cms-pog-conveners-jetmet@cern.ch 2009/08/07

Performance of Jet Algorithms in CMS

The CMS Collaboration

Abstract

The CMS Combined Software and Analysis Challenge 2007 (CSA07) is well underway and expected to produce a wealth of physics analyses to be applied to the first incoming detector data in 2008. The JetMET group of CMS supports four different jet clustering algorithms for the CSA07 Monte Carlo samples, with two different parameterizations each: Fast k_T, SISCone, Midpoint Cone, and Iterative Cone. We present several studies comparing the performance of these algorithms using QCD dijet and $t\bar{t}$ Monte Carlo samples. We specifically observe that the SISCone algorithm performs equal to or better than the Midpoint Cone algorithm in all presented studies and propose that SISCone be adopted as the preferred cone-based jet clustering algorithm in future CMS physics analyses, as it is preferred by theorists for its infrared- and collinear-safety to all orders of perturbative QCD. We furthermore encourage the use of the Fast k_T algorithm which is found to perform as good as any other algorithm under study, features dramatically reduced execution time w.r.t. previous implementations of the k_T algorithm, and is infrared- and collinear save as well.

1 Introduction

Almost every process of interest at the LHC contains quarks or gluons in the final state. The partons can not be observed directly, but fragment into stable hadrons, which can be detected in the tracking and calorimeter systems. This note describes the latest performance studies of several algorithms which cluster energy deposits in the CMS calorimeters into collimated objects of stable particles, "CaloJets". Calorimeter jets are expected to yield a good description of both the parton-level and the hadron showers emerging from the hard interaction. For Monte Carlo events, the hadron-level is defined by applying the same clustering algorithms, which are typically formulated to accept any set of four vectors as inputs, to all stable particles from the MC truth record ("GenJets"). Hadron-level is also referred to as "particle-level", and jet energy scale corrections based on MC are derived to correct back to this detector-independent level.

Calorimeter jets are reconstructed using energy deposits in calorimeter towers ("CaloTowers") as inputs: they are composed of one or more HCAL cells and corresponding ECAL crystals. The unweighted sum of energy deposits of one single HCAL cell and 5x5 ECAL crystals form a projective tower in the barrel ($|\eta| < 1.4$). A more complex association between HCAL cells and ECAL crystals is required in the forward region. The standard jet reconstruction applies the "Scheme B" thresholds on calorimeter cells and the overall tower threshold $E_{\mathrm{T}} > 0.5\,\mathrm{GeV}$, summarized in Table 1 and relevant for all studies presented in this note.

Scheme	HB [GeV]	HO [GeV]	HE [GeV]	$\sum EB$ [GeV]	$\sum EE$ [GeV]
B	0.90	1.10	1.40	0.20	0.45

Table 1: Energy thresholds (in GeV) for calorimeter noise suppression "Scheme B". $\sum EB$ and $\sum EE$ refer to the sum of ECAL energy deposits associated with the same tower in the barrel and endcap respectively.

The studies presented in this note are based on QCD dijet and $t\bar{t}$ Monte Carlo samples without pileup, produced and reconstructed with CMSSW_1_5_2 and analyzed with CMSSW_1_6_X. It is often necessary to associate CaloJets with GenJets in these samples to probe how well the calorimeter-level reconstruction represents the hadron-level of the process. This association is based on spatial separation in η-ϕ-space between the two jet axis by requiring

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} \tag{1}$$

to be less than a certain value. Similarly, GenJets (and hence their associated CaloJets) are assigned the same parton flavor as the matched MC parton from the hard interaction.

Besides good correspondence to the parton-level and hadron-level, a successful jet algorithm should fulfill two important requirements. It should be *collinear-safe*, such that the outcome remains unchanged if e.g. the energy carried by a single particle is instead distributed among two collinear particles. Collinear safety is typically endangered if the jet finding is based on energetic seeds and a threshold is applied to these seeds. The algorithm should be *infrared-safe*, such that the result of the jet finding is stable against the addition of soft particles. Jet algorithms which don't comply with either or both of these requirements yield ambiguous results and lead to unnecessary uncertainties when applied to calculations in perturbative theory.

The performance of the following four jet clustering algorithms that are supported by the CMS JetMET group for CSA07 samples are discussed in this note:

- The Iterative Cone algorithm is a simple cone-based algorithm employed by CMS online in the High Level Trigger (HLT). It has a short and predictable execution time.

Calorimeter towers or particles with $E_T > 1$ GeV are considered in descending order as starting points (seeds) for an iterative search for stable cones such that all inputs with $\sqrt{\Delta\eta^2 + \Delta\phi^2} \leq R$ from the cone axis are associated with the jet, R being the cone size parameter. A cone is considered stable if its geometric center agrees with the (η, ϕ) location of the sum of the constituent four vectors within a certain tolerance. Once a stable cone is found, it is declared a jet and its constituents are removed from the remaining inputs. The algorithm is neither collinear- nor infrared-safe.

- The Midpoint Cone [1] algorithm is based on an iterative procedure to find stable cones as well. Infrared-safety is addressed however by considering the midpoints between each pair of (proto-)jets which are closer than twice the cone radius R as additional seeds. Moreover, each input can initially be associated with several protojets, and a splitting and merging algorithm is applied afterwards to ensure each input appears in one jet only. Despite these improvements to the cone-based clustering procedure, the algorithm has been shown not to be infrared-safe for pQCD orders beyond NLO. Note that the same seed requirements as for the Iterative Cone algorithm are imposed.
- SISCone [2] is the "Seedless Infrared-Safe Cone" jet algorithm. It is collinear- and infrared-safe to all orders of pQCD and demands only slightly higher execution time compared to the Midpoint Cone algorithm. The code is supported and available publicly with a detector-independent interface ensuring that different experiments can compare results with the exact same clustering code applied.
- Fast $\boldsymbol{k_T}$ [3] is a recent implementation of the k_T algorithm[4] which is also collinear- and infrared-safe. It has a dramatically reduced execution time w.r.t. previous implementations of the k_T algorithm. It is the only sequential recombination jet algorithm currently supported in CMS. The radius parameter D plays the corresponding role as the cone size parameter R for cone algorithms, and by construction any pair of clustered k_T jets is guaranteed to be separated by $\sqrt{\Delta y^2 + \Delta\phi^2} > D$.

These four algorithms can be grouped into two general categories: seeded and seedless. The "E-Scheme" is used for all algorithms as the recombination scheme: the energy and momentum of a jet are defined as the sums of energies and momenta of its constituents.

Figure 1 shows the CPU time requirements for each algorithm to cluster all calorimeter towers in the event passing the E_T thresholds, using QCD dijet events. The execution time of the Fast k_T algorithm is comparable to the Iterative Cone algorithm without the discussed deficiencies of the latter. The SISCone algorithm requires more CPU resources compared to the Midpoint Cone algorithm. The time spent for the jet reconstruction ($\approx 0.02\,s$) of each event however is small compared to the total event reconstruction time ($\approx 10\,s$): the particular jet algorithm choice does therefore not impact the overall CPU requirements per event significantly. A version of SISCone with execution times reduced by about 10% was recently made available and will be adopted by CMS in the near future.

The following section summarizes the results of various performance studies and comparisons of the algorithms introduced above.

2 Summary of Jet Performance Studies

In this Section we summarize the results of performance studies comparing jets reconstructed with the algorithms and respective radius parameter (R/D) choices currently supported for CMS analysis. The performance of the CMS calorimeters is known to be different in the barrel

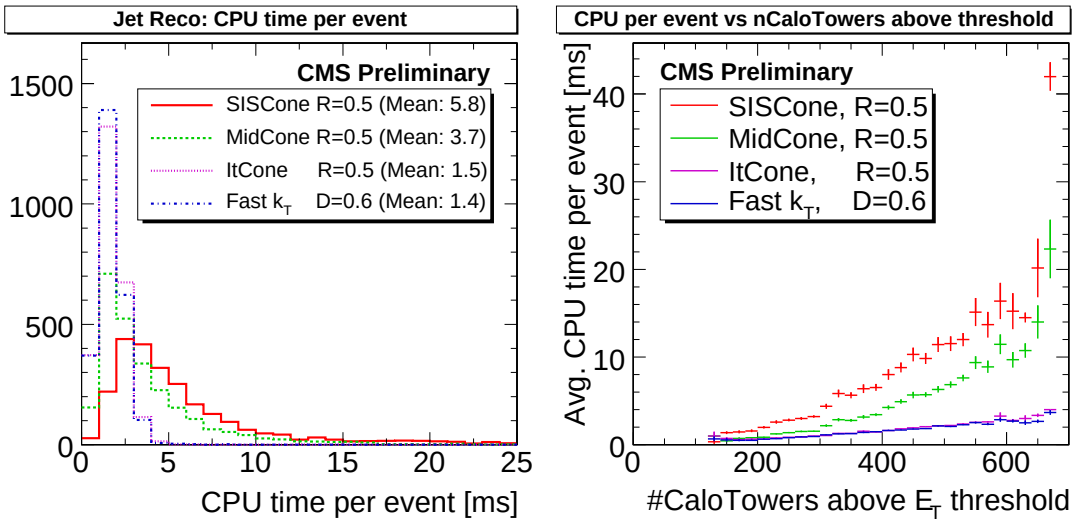

Figure 1: *Left:* CPU time required for each jet algorithm to cluster all CaloTowers above the E_{T} threshold of 0.5 GeV into jets. *Right:* Average CPU time as a function of the number of CaloTowers above E_{T} threshold.

($|\eta| < 1.4$), endcaps ($1.4 < |\eta| < 3$), and forward ($3 < |\eta| < 5$) regions. Many studies presented here are therefore carried out for each of the regions separately, and significant differences are indeed observed. In this note we are however mostly concerned with the relative performance between different algorithms and radius parameter choices, in our quest to select a set of algorithms to be supported for future CMS analyses. The relative performance between different algorithms appears to be consistent across all regions of the detector, and only distributions for the barrel region are therefore shown, while the differences observed in other regions are explained in the text.

The **jet matching efficiency** is defined as the ratio of the number of particle jets matched to a calorimeter jet within $\Delta R < 0.5$ and the total number of particle jets. It represents a meaningful measure of the reconstruction efficiency of each jet algorithm, but is strongly correlated to the position resolution and therefore depends on the ΔR cut and the jet size parameter. However, relative comparisons between different algorithms using equivalent size parameters remain instructive. The matching efficiencies for small (left) and large (right) radius parameters as a function of $p_{\mathrm{T}}^{\mathrm{gen}}$ are shown in Figure 2. The efficiencies of jets reconstructed with the Fast k_{T} and SISCone algorithms indicate better performance than jets reconstructed with the Midpoint Cone and Iterative Cone algorithms.

The **jet response** $\mathrm{R}_{jet} = p_{\mathrm{T}}/p_{\mathrm{T}}^{\mathrm{gen}}$ for the barrel region as a function of $p_{\mathrm{T}}^{\mathrm{gen}}$ is shown in Figure 3 for uncorrected jets. The results for small radius parameters ($R = 0.5/D = 0.4$) are shown on the left side of each of the Figures, large radius parameters ($R = 0.7/D = 0.6$) are covered on the right side. Very good agreement between the individual algorithms is found for all regions of the detector, indicating good correspondence between the values of D for the k_{T} algorithm and R for cone algorithms which are being compared.

The η **and** ϕ **resolutions** for jets in the barrel region are shown as a function of $p_{\mathrm{T}}^{\mathrm{gen}}$ in Figures 4 and 5 respectively. Good agreement is found among all algorithms with comparable radius parameter, with marginal differences at low $p_{\mathrm{T}}^{\mathrm{gen}}$. Jets reconstructed with larger radius parameters yield slightly worse resolution both in η and ϕ. Note that the position of the primary vertex is assumed to be at $z = 0$, which dilutes the η resolution w.r.t. taking the correct position measured with the tracking detectors into account.

Figure 6 shows the **jet energy resolutions** derived from MC truth for jets in the barrel region. Jets reconstructed with Fast k_T show slightly worse resolution at low p_T^{gen}, while no significant impact of the radius parameter choice is observed. The resolutions are obtained additionally without using MC truth information by using the data-driven *Asymmetry Method*, which relates the jet p_T resolution to the resolution of the p_T-imbalance between the two leading jets. A soft radiation correction is derived by selecting events with an additional 3^{rd} jet and studying the measured resolution as a function of various maximum p_T cuts on the extra jet (illustrated in Figure 7 on the left). The results are compared to the MC truth derived resolutions on the right of Figure 7 for jets reconstructed with Iterative Cone $R = 0.5$. Good agreement is observed for the p_T region studied, demonstrating that the jet energy resolution can be extracted from dijet events.

The jet reconstruction **performance in $t\bar{t}$ events** is studied by selecting events with one ("lepton+jets") or zero ("alljets") electron or muon in the final state from a $t\bar{t}$ ALPGEN sample with no additional jets ("$t\bar{t}$ +0 jets"). For each of the four algorithms, an additional even smaller radius parameter choice ($R = 0.4/D = 0.3$) is considered. $t \to bq\bar{q}'$ and $\bar{t} \to \bar{b}\bar{q}q'$ decays are identified on particle level and only events are considered for which all three decay products of one or both $t(\bar{t})$ decay(s) can be uniquely matched to reconstructed calorimeter jets. The efficiency to select these decays indicates the performance of the respective jet algorithm in a busy multijet environment and its ability to correctly resolve the topology of the underlying process. The Fast k_T algorithm is hereby found to fully resolve hadronic $t(\bar{t})$ on calorimeter level more efficiently than any cone-based algorithm. For the selected events, the invariant two-jet (W boson) and three-jet (top quark) masses are compared on particle-level ("GEN"), calorimeter-level ("CALO"), corrected calorimeter-level ("CORR"), and corrected calorimeter-level with additional flavor-dependent corrections applied ("Level 5" or "L5"). The m_W and m_t distributions obtained for all correction levels are shown in Figure 8 for jets reconstructed with Fast k_T $D = 0.4$. Figure 9 shows the comparison of the obtained relative widths (RMS$(m)/m$) of the m_W (top) and m_t (bottom) distributions. The obtained mass resolutions are in good agreement for all algorithms and radius parameters, with the exception that the mass resolution improves slightly for jets reconstructed with Fast k_T for larger radius parameters D.

The internal properties of a jet are studied for all algorithms by comparing the multiplicities and p_T distributions of the constituents of both particle jets (MC particles) and calorimeter jets (calorimeter tower energy deposits). Both particle- and calorimeter-level distributions are in good agreement. Multiplicities increase logarithmically with the transverse momentum of the jet, and the p_T distribution of the inputs becomes harder for higher jet momenta. The internal structure of the jet can also be described by the **energy distribution within a jet** which is characterized by the differential jet shape, $\rho(r)$, defined as

$$\rho(r) = \frac{\sum p_T(r - \Delta r/2, r + \Delta r/2)}{\Delta r \sum p_T^{Jet}}$$

where the sum is over all the jet constituents in the range $(r - \Delta r/2, r + \Delta r/2)$ in the numerator and $r = \sqrt{(y_{jet} - y_c)^2 + (\phi_{jet} - \phi_c)^2}$ with (y_{jet}, ϕ_{jet}) and (y_c, ϕ_c) being the position of the jet and the constituents. The denominator $\sum p_T^{Jet}$ is the scalar sum of the transverse momenta of all the jet constituents. The jet shapes for particle and calorimeter jets in the range $80 < p_T < 120\,\text{GeV}$ are shown in Figure 10 for particle jets (left) and calorimeter jets (right) respectively. Jets become narrower with increasing jet p_T. Note that the Iterative Cone algorithm is based on $\Delta R(\eta)$, while the differential jet energy density is defined using $\Delta r(y)$, explaining contributions to the density for $r > R$.

Figure 11 shows the **dijet mass resolution** as a function of the resonance mass $m_{Z'}$ for jets reconstructed with Midpoint Cone (dashed red line) and SISCone (solid blue line). The Z' Monte Carlo sample is intentionally miscalibrated according to the expectation of the quality of the calibration of the CMS detector after 100 pb^{-1} of data taking. The dijet mass is computed as the invariant mass of the two leading jets in events where both leading jets are reconstructed in the barrel ($|\eta| < 1.3$). The individual resolutions are obtained from a Gaussian fit to each distribution in the range -1σ to 1.5σ centered on the mean. The mass resolutions achieved with both algorithms are in good agreement over the entire studied range of resonance masses.

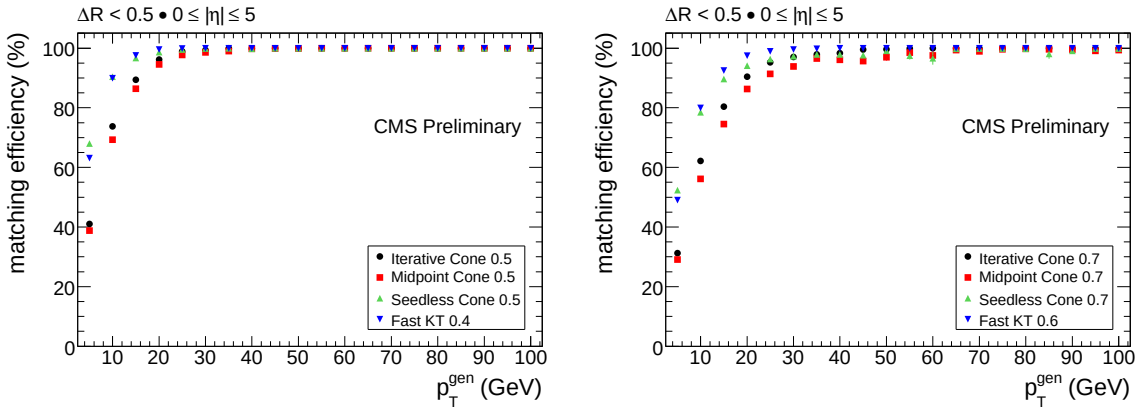

Figure 2: Matching Efficiency vs $p_{\mathrm{T}}^{\mathrm{gen}}$ for $R = 0.5/D = 0.4$ (left) and $R = 0.7/D = 0.6$ (right) jets.

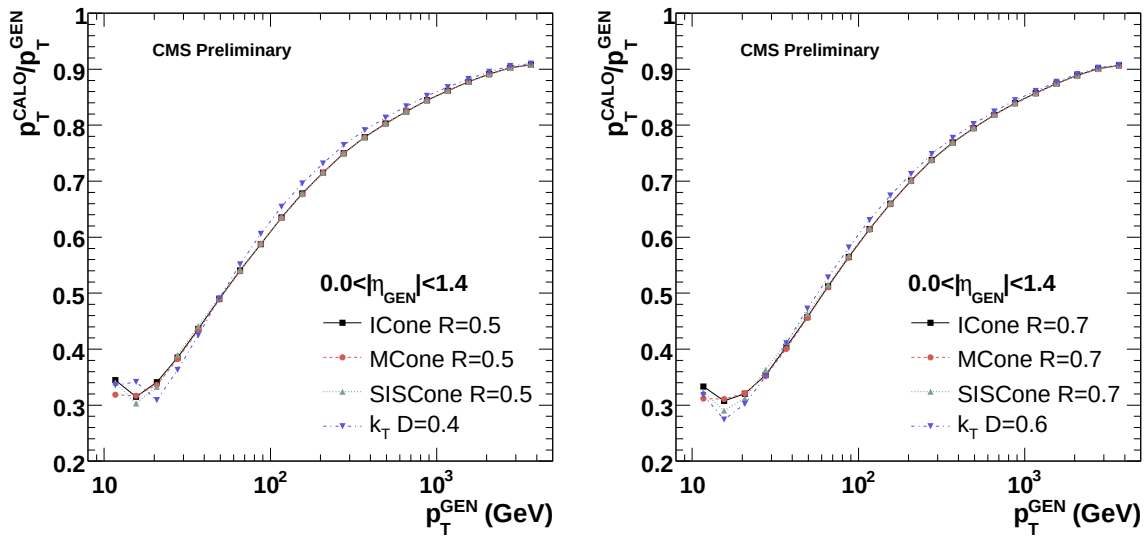

Figure 3: The jet response as a function of $p_{\mathrm{T}}^{\mathrm{gen}}$, averaged over the Barrel region, for jets clustered with smaller (left) and larger (right) size parameters.

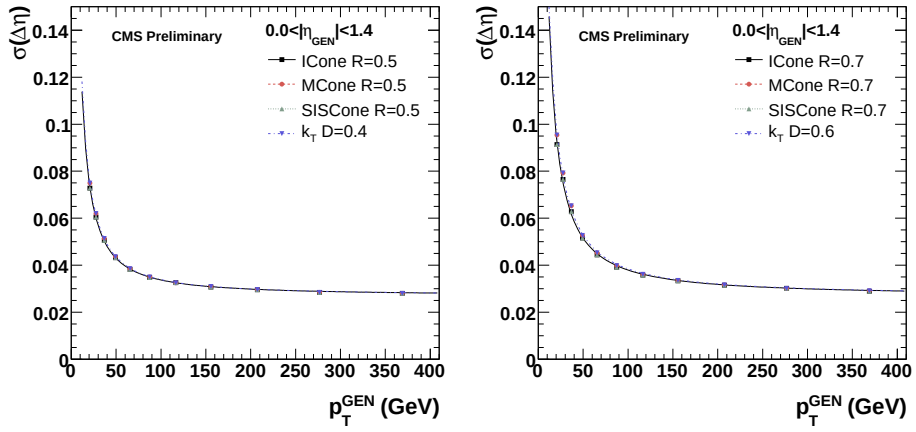

Figure 4: The jet η resolutions as a function of $p_{\mathrm{T}}^{\mathrm{gen}}$, averaged over the Barrel region, for jets clustered with smaller (left) and larger (right) size parameters. The resolutions are derived using MC truth information.

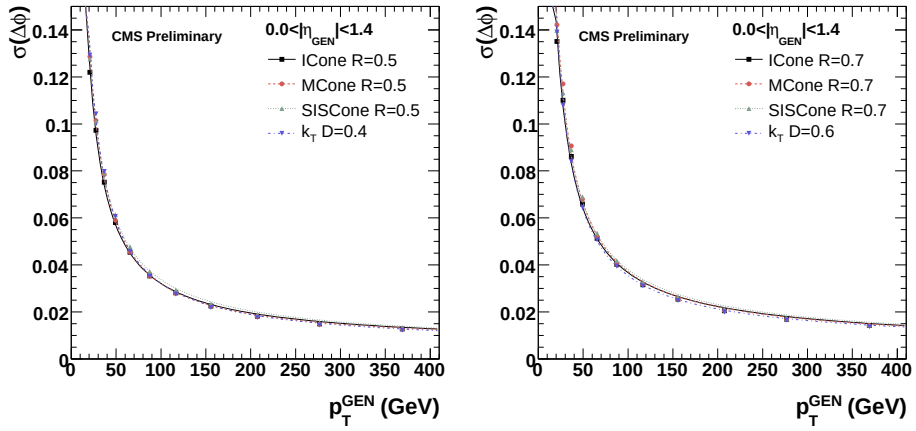

Figure 5: The jet ϕ resolutions as a function of $p_{\mathrm{T}}^{\mathrm{gen}}$, averaged over the Barrel region, for jets clustered with smaller (left) and larger (right) size parameters. The resolutions are derived using MC truth information.

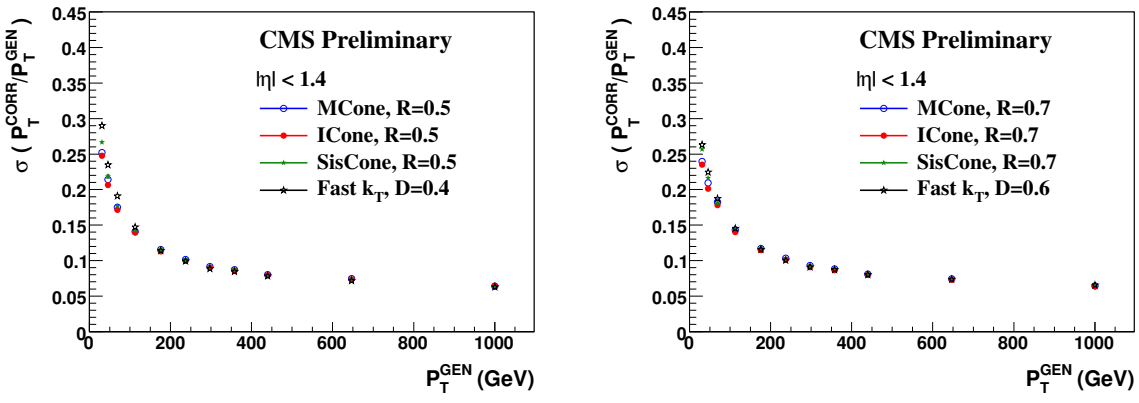

Figure 6: Jet energy resolution derived from MC truth for Midpoint Cone, Iterative Cone, SIS-Cone, and Fast k_{T} with $R = 0.5/D = 0.4$ (left) and $R = 0.7/D = 0.6$ (right) in the barrel region ($|\eta| < 1.4$).

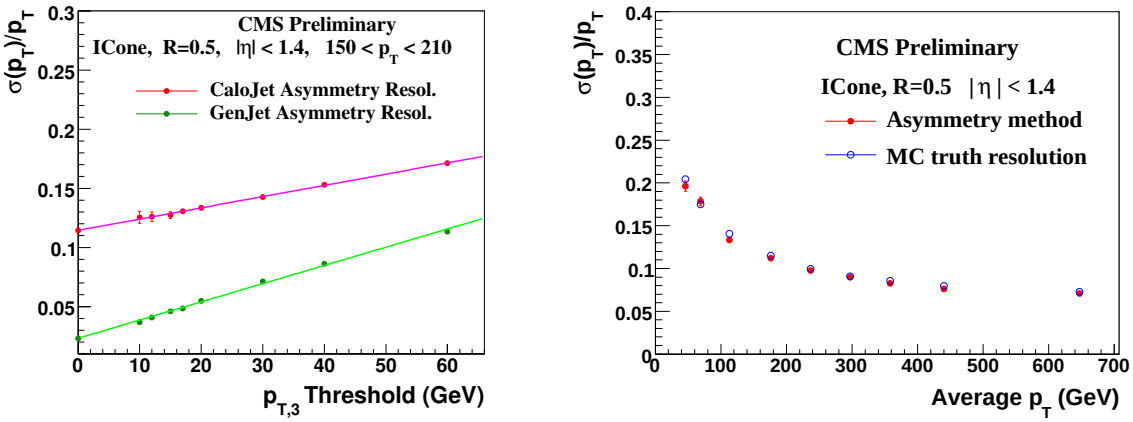

Figure 7: *Left:* Resolution from dijet asymmetry as a function of the p_{T} threshold applied to the third jet in the event for Iterative Cone $R = 0.5$, extrapolated to zero. The red and green lines correspond to detector and particle level respectively. The plots are taken from the average p_{T} bin with $150 < p_{\mathrm{T}} < 210\,\mathrm{GeV}$. *Right:* Resolution obtained with the Asymmetry Method and from Monte Carlo Truth for Iterative Cone $R = 0.5$ jets.

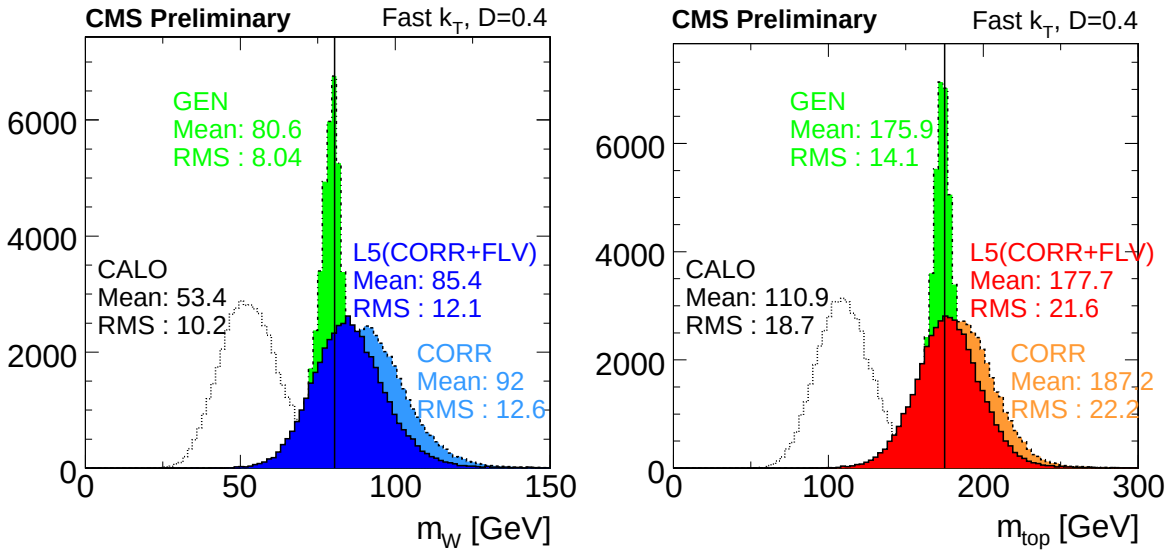

Figure 8: m_W and m_t distributions for hadronic top decays reconstructed with the Fast k_{T} algorithm, $D = 0.4$. Distributions are shown for particle-level jets (GEN), calorimeter jets (CALO), calorimeter jets corrected with MCJet corrections (CORR), and corrected calorimeter jets with an additional flavor correction ("Level-5 correction") applied (L5). Only jets with uncorrected $p_{\mathrm{T}} \geq 15\,\mathrm{GeV}$ and $|\eta| \leq 5$ are considered. The generated W boson (80.42 GeV) and top quark (175 GeV) masses are indicated by the black vertical lines.

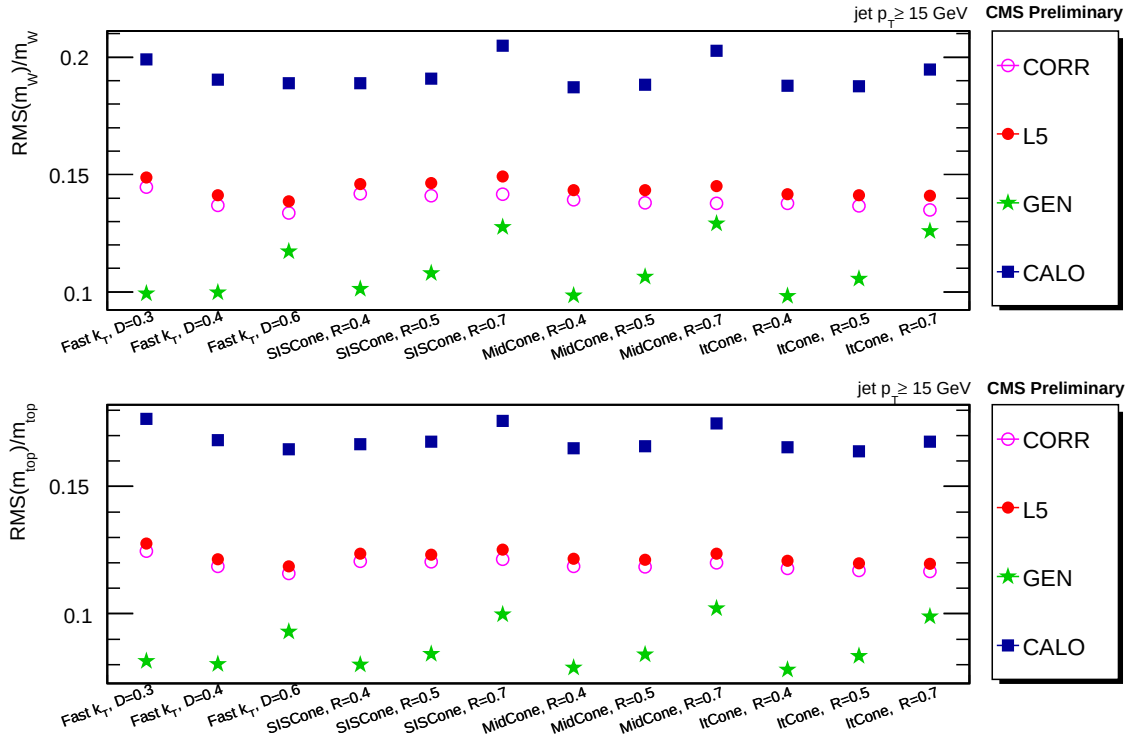

Figure 9: Relative width (RMS$(m)/m$) of m_W (top) and m_t (bottom) distributions for all studied jet algorithms for particle-level jets (GEN), calorimeter jets (CALO), corrected calorimeter jets (CORR), and corrected calorimeter jets with an additional flavor correction applied (L5). Only hadronic $t\bar{t}$ decays fully matched to three calorimeter jets with uncorrected $p_{\mathrm{T}} \geq 15\,\mathrm{GeV}$ and $|\eta| \leq 5$ are considered.

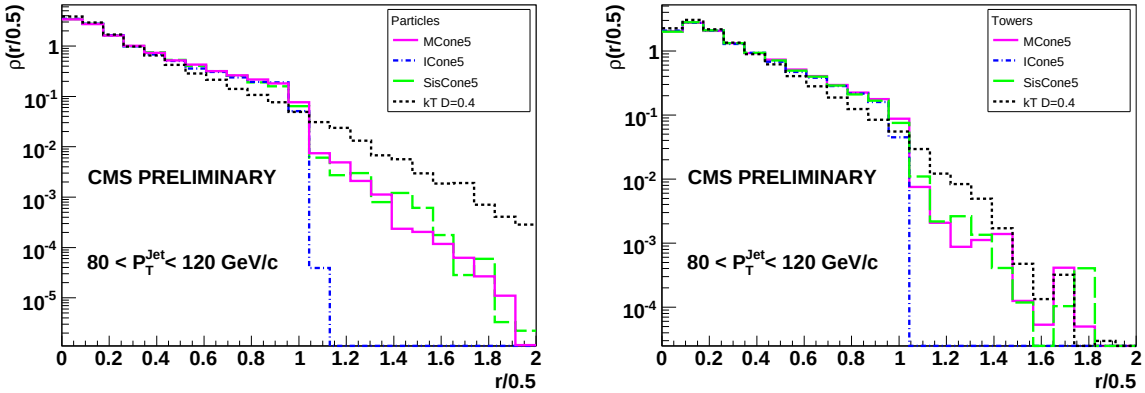

Figure 10: Normalized transverse energy density distributions ($\rho(r)$) in particle jets (left) and calorimeter jets (right) in a particular p_{T} range ($80 < p_{\mathrm{T}} < 120\,\mathrm{GeV}$) for small radius parameters $R = 0.5/D = 0.4$.

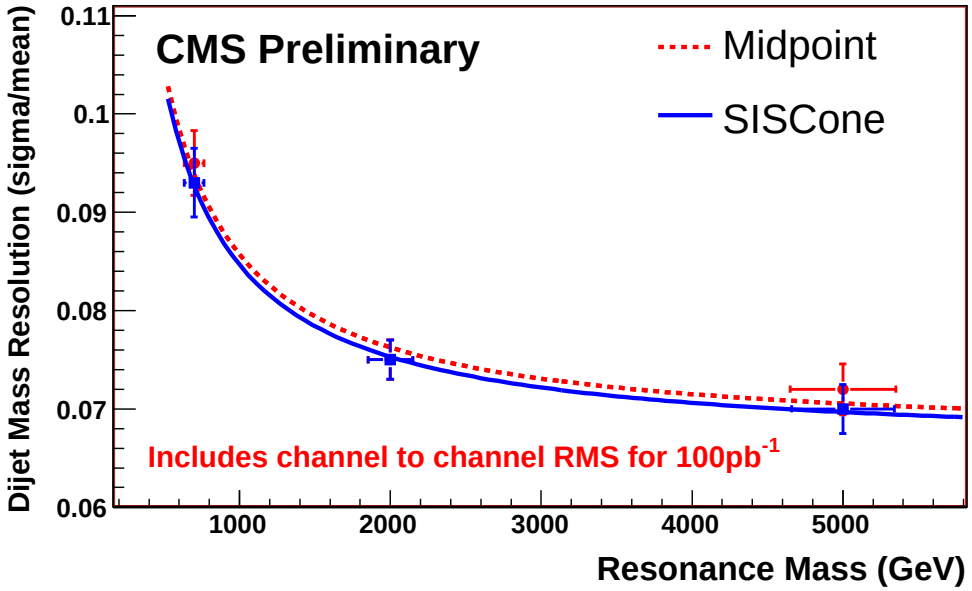

Figure 11: Comparison of the dijet mass resolution as a function of the resonance mass $m_{Z'}$ for jets reconstructed with Midpoint Cone (dashed red line) and SISCone (solid blue line). The cone size parameter is $R = 0.5$ in both cases.

3 Conclusions

We presented detailed comparisons of performance between four jet reconstruction algorithms currently available in CMSSW with two radius parameter choices each: Iterative Cone, SISCone, and Midpoint Cone with $R = 0.5$ and 0.7, and Fast k_{T} with $D = 0.4$ and 0.6. The performance comparisons presented in this note include jet energy response, position resolutions, energy resolutions, efficiencies in QCD dijet samples, reconstruction of the more complex $t\bar{t}$ signal, jet composition and shape distributions, and dijet mass resolution in Z' events. We have developed two data-based techniques to derive the jet energy resolution, which agree well with results based on MC truth for $p_{\mathrm{T}} > 300\,\mathrm{GeV}$ and are within 10% for lower momenta.

We find similar performance on the calorimeter level between algorithms with similar size parameter. The impact of the detector effects appears to be more pronounced than the algorithmic differences studied in this note. We also find that the SISCone algorithm performs as well or better than the Midpoint Cone, while known to be preferred theoretically. Therefore we recommend to adopt SISCone as the default cone-based jet algorithm and consequently to include it in the reconstruction in future standard event processing at CMS.

The k_{T} algorithm is infrared- and collinear safe to all orders of pQCD as well and complementary to the cone-based algorithms. The execution time of Fast k_{T} is dramatically reduced w.r.t. earlier k_{T} implementations and it is therefore well suited for the high multiplicity environment of LHC pp collisions, in fact executing faster than all cone-based algorithms but Iterative Cone. We find that it performs as good or better than any other algorithm in this note and strongly encourage its use as an alternative to SISCone. Further studies will be conducted regarding the performance of all algorithms in events with high pileup and more realistic calorimeter noise.

References

[1] G. C. Blazey *et al.*, *"Run II jet physics: Proceedings of the Run II QCD and Weak Boson Physics Workshop"*, hep-ex 0005012 (2000).

[2] G. P. Salam and G. Soyez, *"A practical seedless infrared-safe cone jet algorithm"*, JHEP05(2007)086 (2007).

[3] M. Cacciari and G. P. Salam, *"Dispelling the N^3 myth for the Kt jet-finder"*, Phys.Lett. B**641** 57-61 (2006).

[4] S. Catani, Y. L. Dokshitzer, M. H. Seymour and B. R. Webber, *"Longitudinally invariant Kt clustering algorithms for hadron hadron collisions"*, Nucl.Phys.B**406**:187-224 (1993).